mall_customer_segmentation

Róbert Vámosi

2022 05 02

R Markdown

This is an R markdawn which shows an easy clusterization of mall customers. The dataset was downloaded from here: https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python?datasetId=42674&language=R

```
mall <- read.csv("Mall_Customers.csv", encoding = ’UTF-8’)
summary(mall)
```

```
##    CustomerID        Gender               Age        Annual.Income..k..
##  Min.   :  1.00   Length:200         Min.   :18.00   Min.   : 15.00  
##  1st Qu.: 50.75   Class :character   1st Qu.:28.75   1st Qu.: 41.50  
##  Median :100.50   Mode  :character   Median :36.00   Median : 61.50  
##  Mean   :100.50                      Mean   :38.85   Mean   : 60.56  
##  3rd Qu.:150.25                      3rd Qu.:49.00   3rd Qu.: 78.00  
##  Max.   :200.00                      Max.   :70.00   Max.   :137.00  
##  Spending.Score..1.100.
##  Min.   : 1.00         
##  1st Qu.:34.75         
##  Median :50.00         
##  Mean   :50.20         
##  3rd Qu.:73.00         
##  Max.   :99.00
```

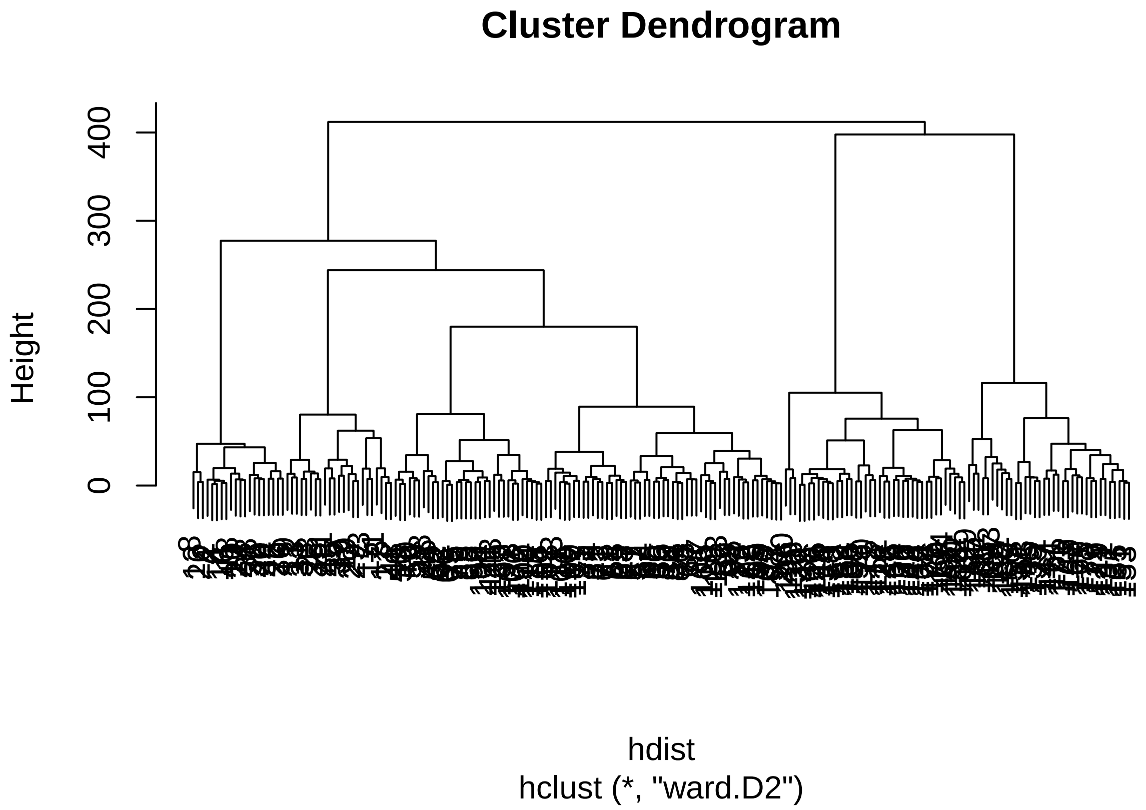
Firstly I change gender variable to a numeric format where Male is 1 and Female is 0.

```
mall<-mall[,-1]

mall.norm<-mall%>%
  mutate(
    Gender = ifelse(Gender == ’Male’,1,0),
  )
```

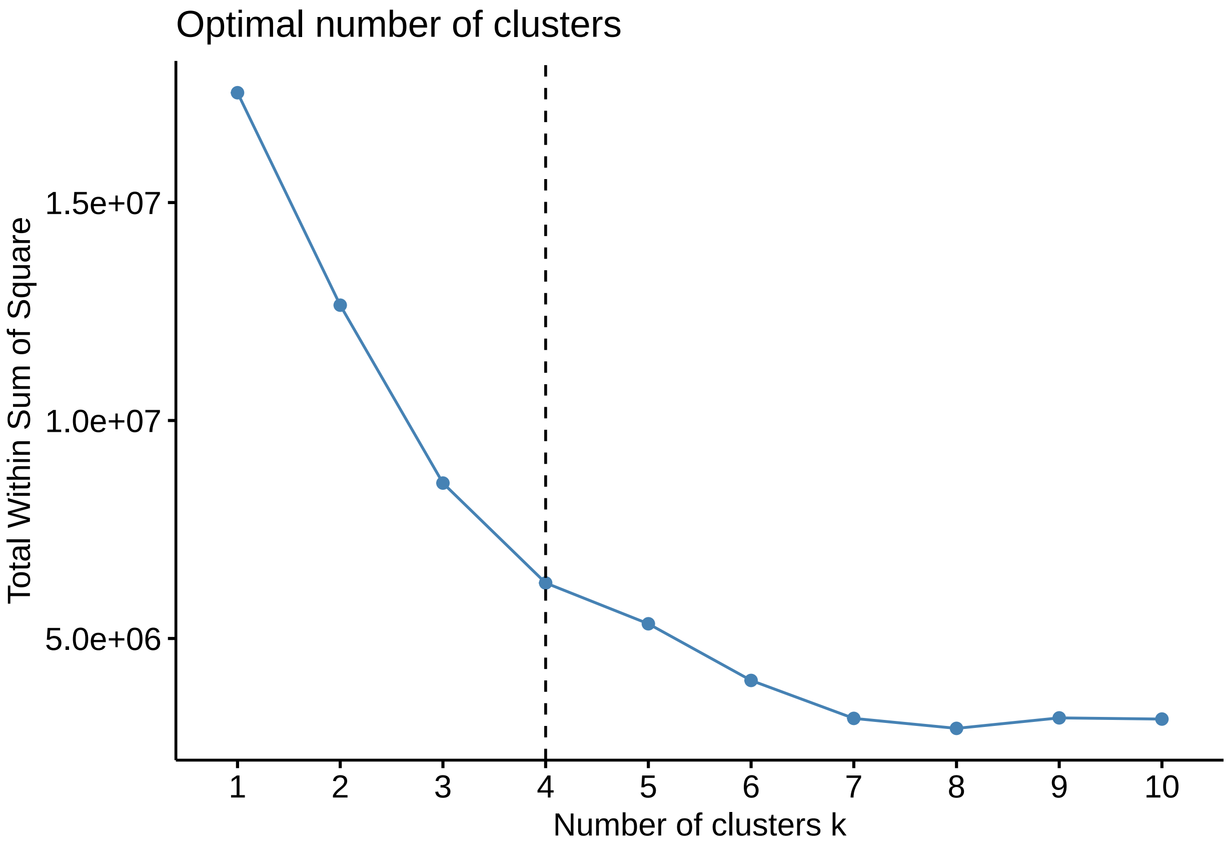
Check a dendogram:

```
set.seed(100)
hdist<-dist(mall.norm) #euclidean method
hclust(hdist, method=’ward.D2’)%>%plot()
```

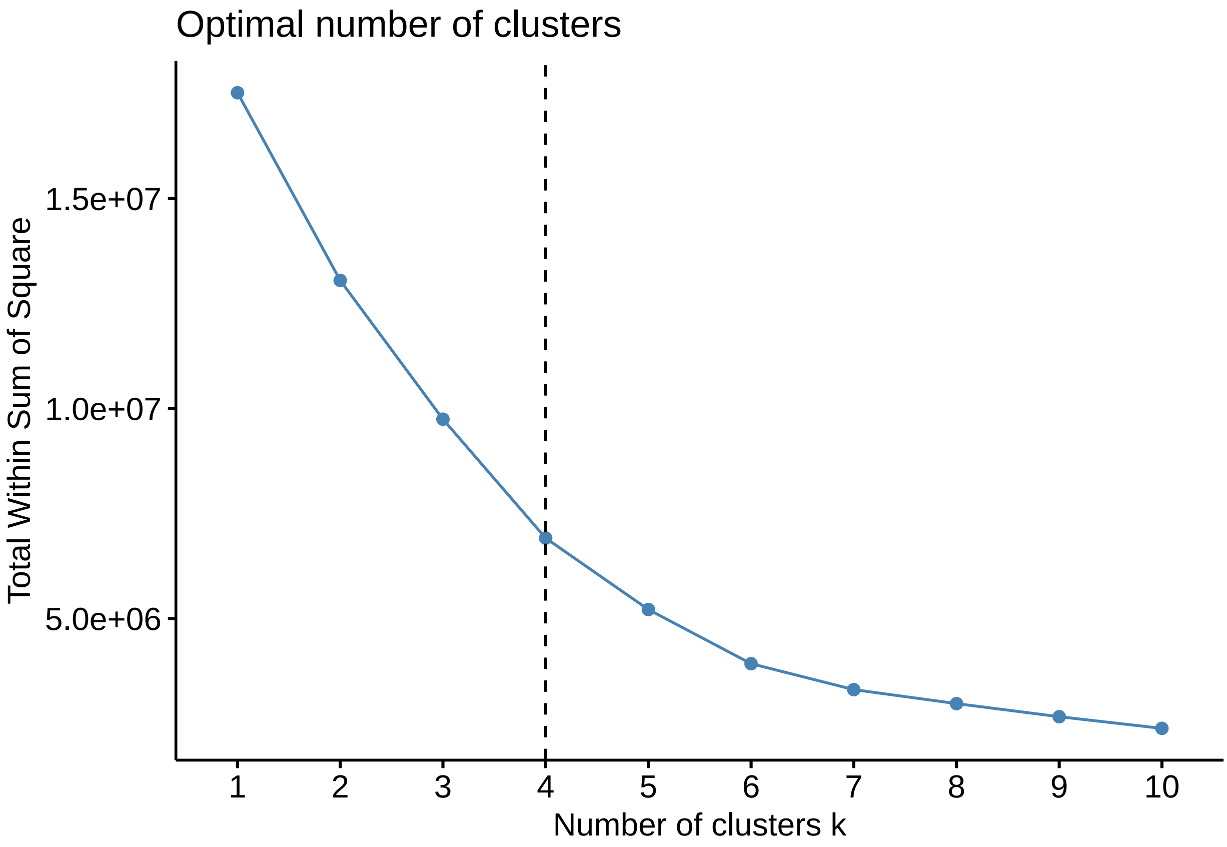


My choice is 4 K depends on these 2 plot.

```
fviz_nbclust(as.matrix(hdist),
             kmeans,
             method = "wss") +
  geom_vline(xintercept = 4, linetype = 2)
```

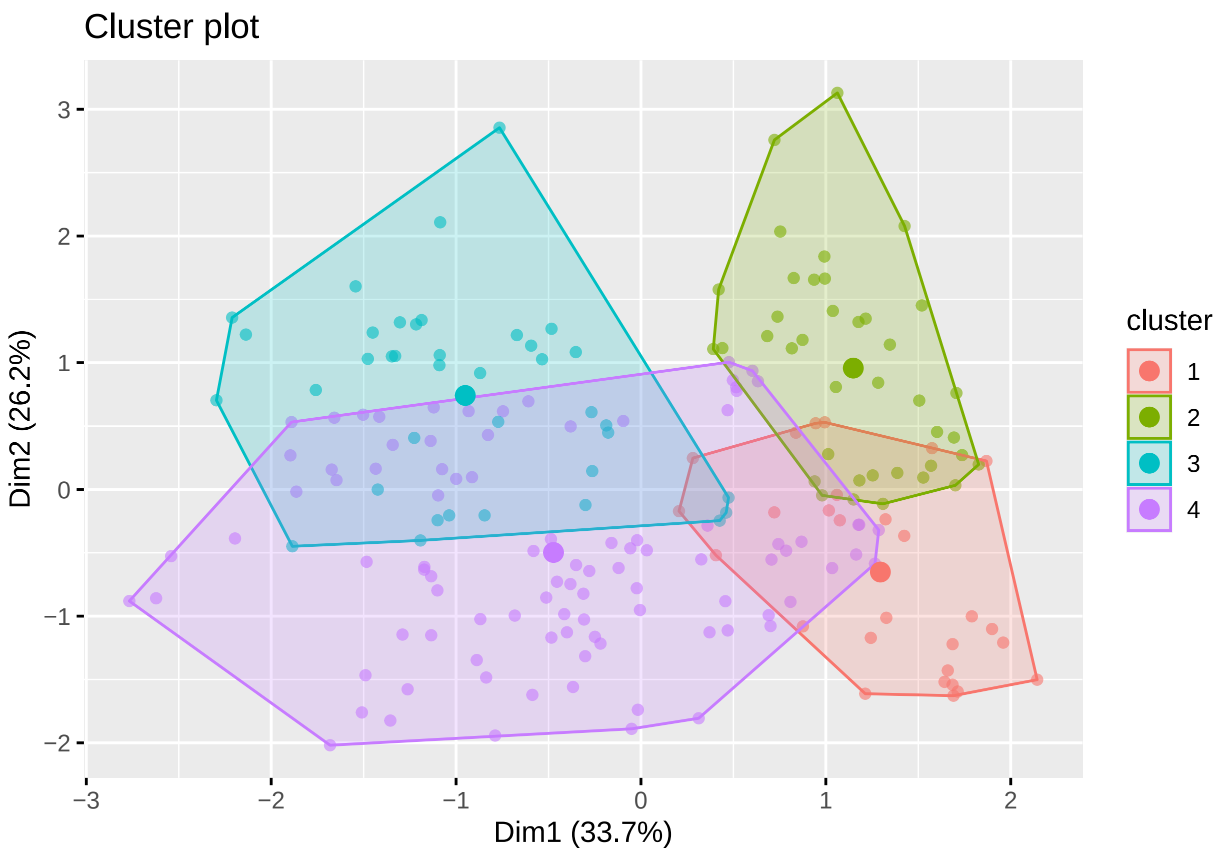


```r
fviz_nbclust(as.matrix(hdist),
             hcut,
             method = "wss") +
  geom_vline(xintercept = 4, linetype = 2)
```



Plot the clusters:

```
kplot<-kmeans(mall.norm, nstart = 25, centers = 4)
fviz_cluster(kplot, data = mall.norm, alpha=0.6 ,shape=19, geom = "point")
```



Create a new data frame which includes group variable

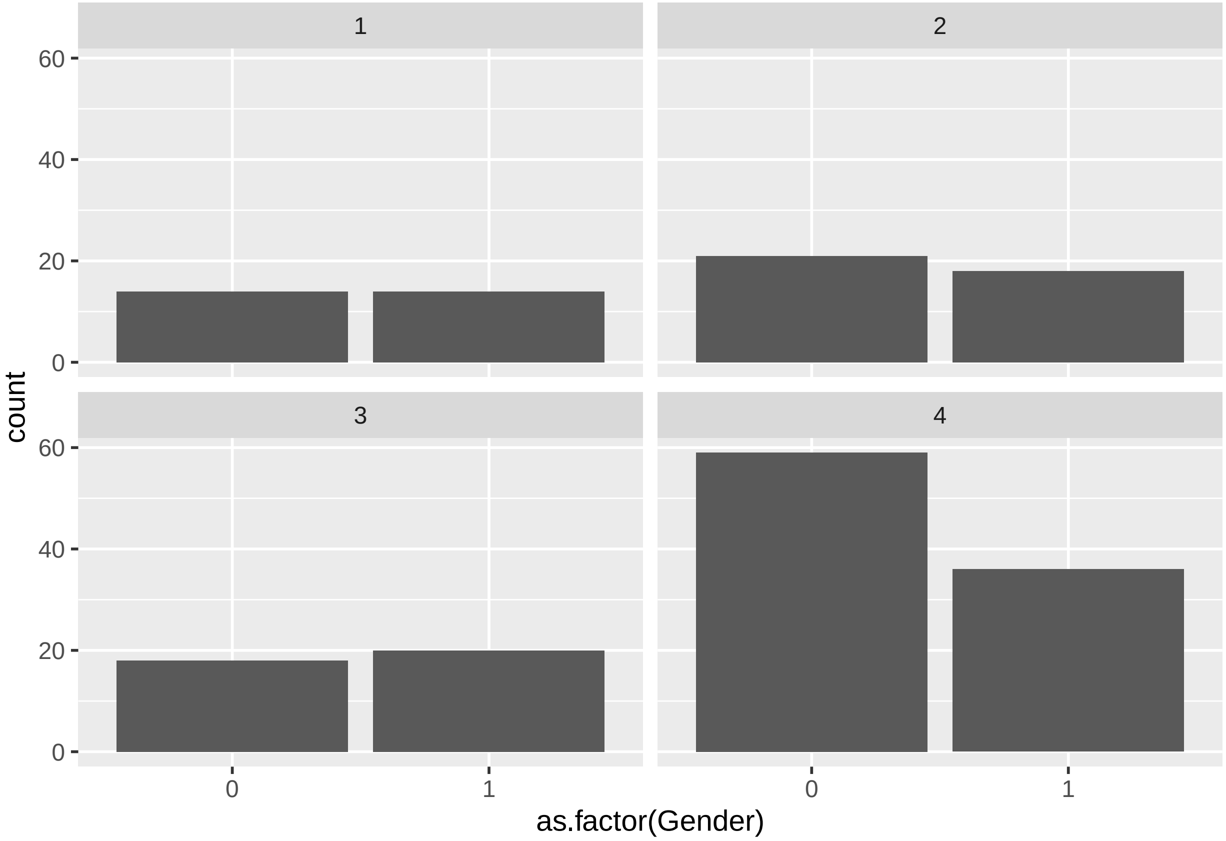
```
mall.new<-mall.norm%>%
  mutate(group=kplot$cluster)

summary(mall.new)
```

```
##      Gender          Age        Annual.Income..k.. Spending.Score..1.100.
##  Min.   :0.00   Min.   :18.00   Min.   : 15.00     Min.   : 1.00
##  1st Qu.:0.00   1st Qu.:28.75   1st Qu.: 41.50     1st Qu.:34.75
##  Median :0.00   Median :36.00   Median : 61.50     Median :50.00
##  Mean   :0.44   Mean   :38.85   Mean   : 60.56     Mean   :50.20
##  3rd Qu.:1.00   3rd Qu.:49.00   3rd Qu.: 78.00     3rd Qu.:73.00
##  Max.   :1.00   Max.   :70.00   Max.   :137.00     Max.   :99.00
##      group
##  Min.   :1
##  1st Qu.:2
##  Median :3
##  Mean   :3
##  3rd Qu.:4
##  Max.   :4
```

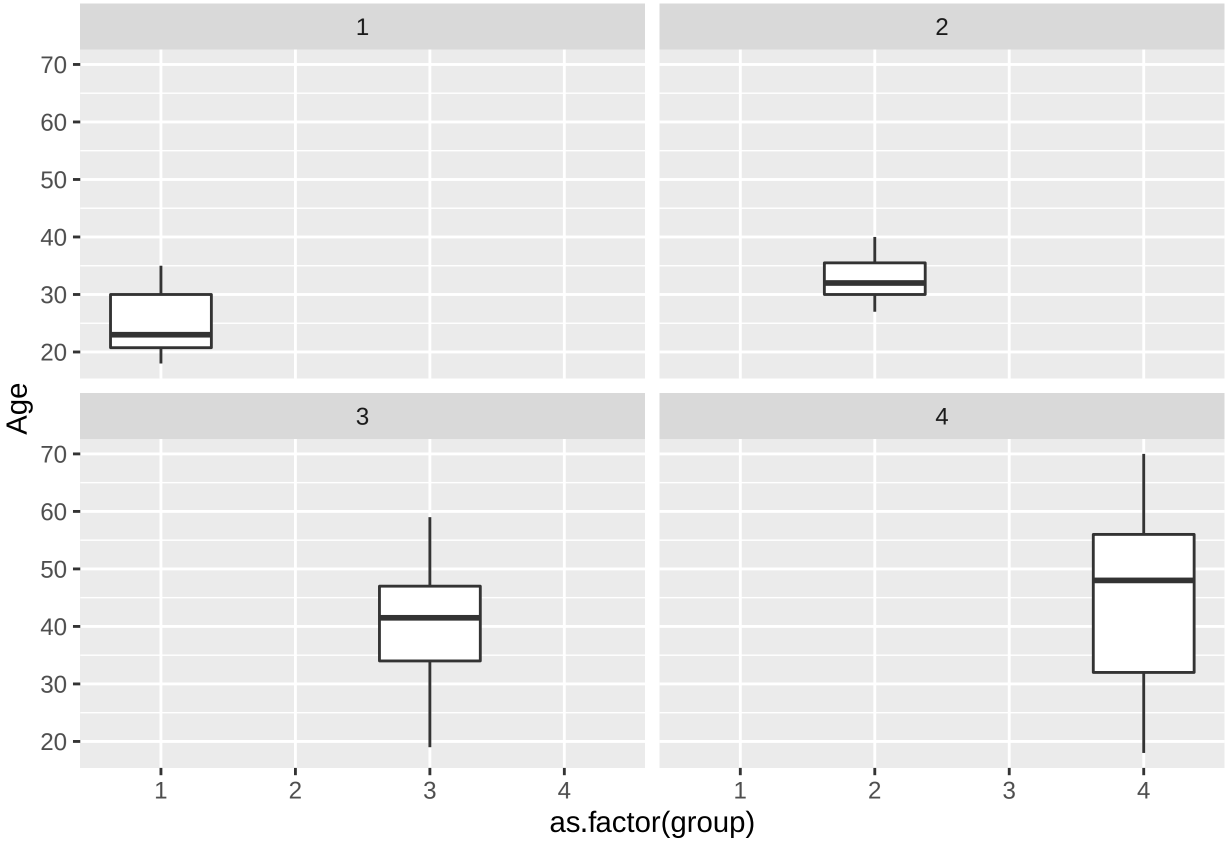
There are more females in group 3 but we cannot observe any difference in the other groups.

```r
mall.new %>% ggplot(aes(as.factor(Gender))) + geom_bar() + facet_wrap(~ group)
```



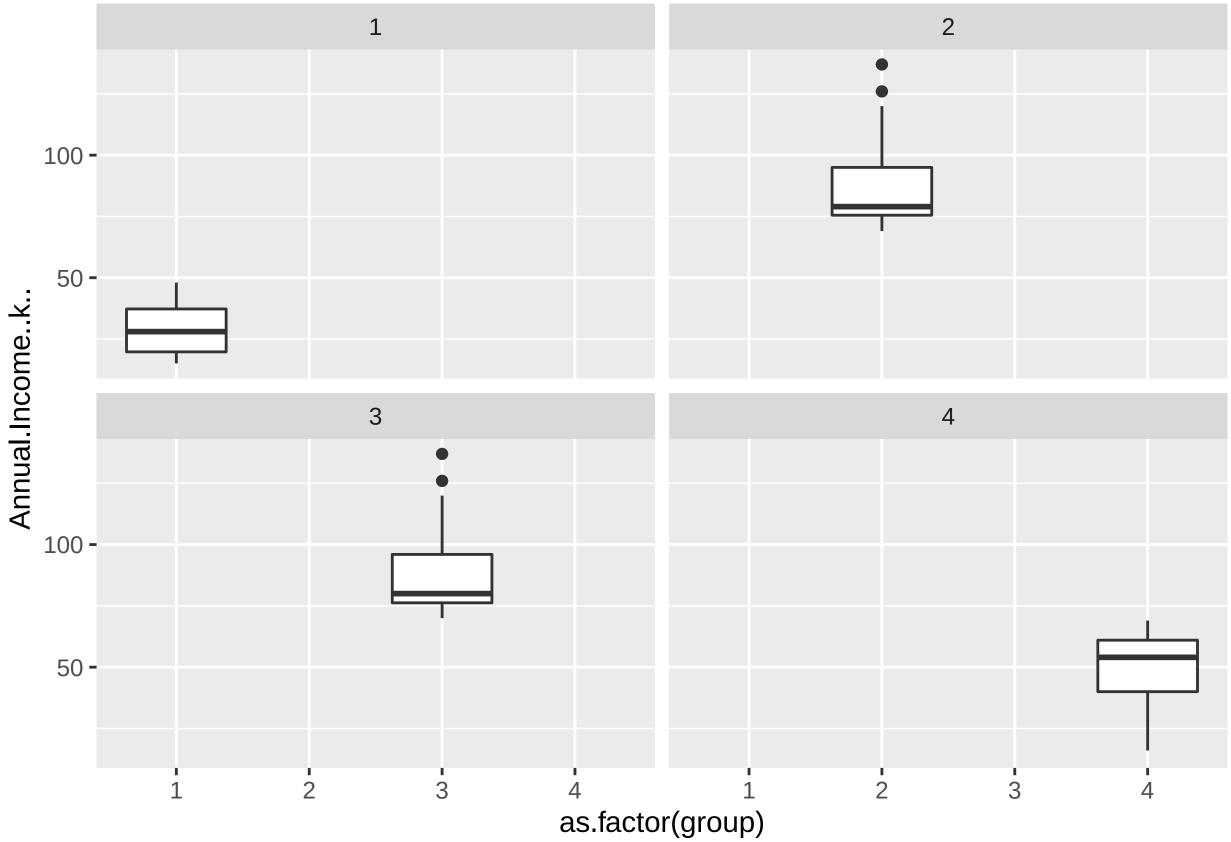
Group 1 and group 3 are older and group 2 and group 4 are younger.

```r
mall.new %>% ggplot(aes(as.factor(group), Age)) + geom_boxplot() + facet_wrap(~ group)
```



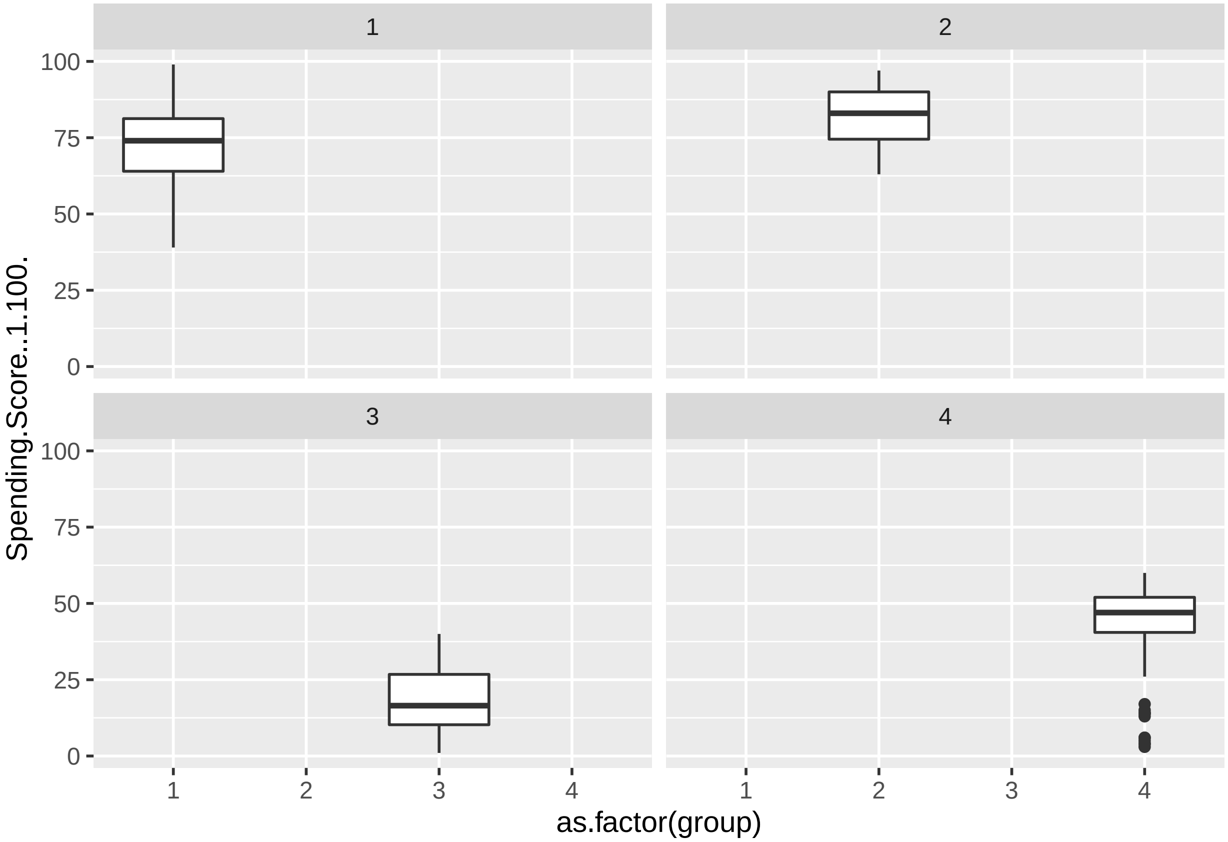
Group 1 and 4 have a higher income and group 2 and 3 have lower.

```
mall.new %>% ggplot(aes(as.factor(group), Annual.Income..k..)) + geom_boxplot() + facet_wrap(~ group)
```



Group 1 and 2 have a higher spending score. Group 4 have a lower spending score.

```
mall.new %>% ggplot(aes(as.factor(group), Spending.Score..1.100.)) + geom_boxplot() + facet_wrap(~ grou
```



For summary: Group 1 - Females and males, the median age is 40, has a higher income but lower spending score. Group 2 - Females and males, the median age is under 30, has a lower income but a higher spending score. Group 3 - Rather females than males, the median age is about 50, has a lower income and lower spending score, but bigger than group 1 Group 4 - Females and males, the median age is about 30, has a higher income and higher spending score.